Roscoff 2013
Formation sRNAseq
Analyse des miRNAs sous Galaxy

- EXERCICES -

Babraham Bioinformatics

"**FastQC** is a quality control tool for high throughput sequence data."
http://www.bioinformatics.bbsrc.ac.uk/

cutadapt

A tool that removes adapter sequences from DNA sequencing reads.
cutadapt removes adapter sequences from high-throughput sequencing data. https://code.google.com/p/cutadapt/

B W A

"**Burrows-Wheeler Aligner** (**BWA**) is an efficient program that aligns relatively short nucleotide sequences against a long reference sequence such as the human genome."
http://bio-bwa.sourceforge.net

SAMtools

"**SAM** (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments."
http://samtools.sourceforge.net

MDC MAX-DELBRÜCK-CENTRUM FÜR MOLEKULARE MEDIZIN BERLIN-BUCH IN DER HELMHOLTZ-GEMEINSCHAFT e.V.

miRDeep2 is a software package for identification of novel and known miRNAs in deep sequencing data. Furthermore, it can be used for miRNA expression profiling across samples. Last, a new module for preprocessing of raw Illumina sequencing data produces files for downstream analysis with the miRDeep2 or quantifier module.

Objectifs :

Cette formation a pour objectif de vous familiariser à l'utilisation des outils dédiés au traitement des données de type sRNAseq sur votre instance Galaxy (sigenae-workbench).

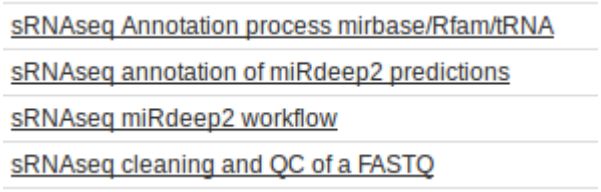
Vous découvrirez notamment comment :

- Contrôler la qualité de vos données brutes
- Nettoyer des données sRNAseq
- Annoter des séquences contre des banques de ncRNA
- Prédire des miRNA avec miRdeep2
- Annoter les prédictions de miRdeep2

Pour vous connecter à l'instance Sigenae de Galaxy :

- http://sigenae-workbench.toulouse.inra.fr
- Si besoin, vous avez la possibilité de demander un compte depuis http://bioinfo.genotoul.fr/index.php?id=74 (avec une adresse académique uniquement).
- Comptes disponibles pour votre session de formation :
  - Logins : anemone, aster, bleuet, iris, muguet, narcisse, pensee, rose, tulipe, violette, lilas, pervenche, laurier, lavande, lis, capucine, coquelicot, geranium, liseron, arome, chardon.
  - Password : f1o2r3!

Les workflows que nous vous demandons de générer sont tous disponibles dans la rubrique «Published Workflows», préfixés par sRNAseq.



Exercice n°1 : Préparation de l'historique Galaxy

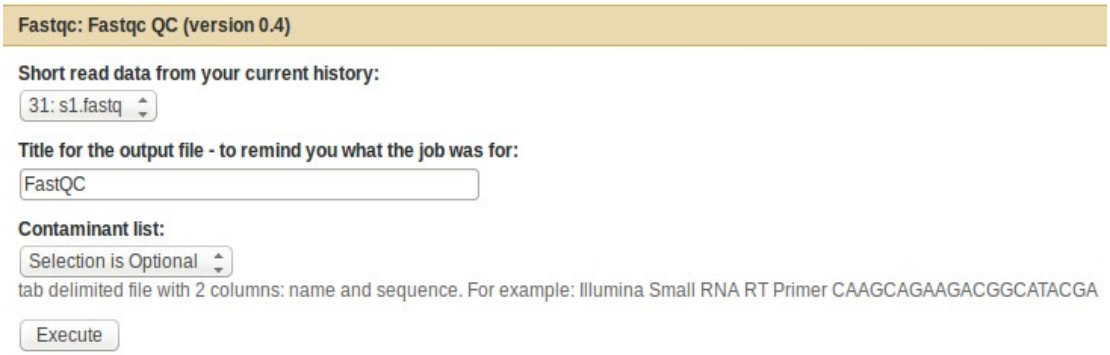
Nous allons travailler à partir de jeux de données Illumina représentant 2 tissus «s2.fastq» et «s1.fastq».
Pour récupérer vos jeux de données, importer cet historique : «**ROSCOFF DATA TP sRNAseq 21/11/2013**».
Cet historique contient les deux datasets **s1.fastq** et **s2.fastq** : ce sont ces fichiers qui serviront pour l'analyse qualité et le nettoyage. Les deux datasets **cleaned_sequences.fasta** et **cleaned_sequences.fastq** sont les deux datasets qui serviront aux exercices de prédiction et d'annotation.

Renommer votre historique.

Exercice n°2 : Analyse de la qualité et nettoyage

1 – Analyse de la qualité des données

A partir du jeu de données s1, utiliser l'outil Galaxy «Fastqc: Fastqc QC using FastQC from Babraham» pour générer les graphiques d'analyse de la qualité de séquençage.



Il vous est possible de récupérer :

- L'ensemble des résultats de FastQC en cliquant sur l'icône «disquette» du dataset.
- Un graphique par un clic droit sur son intitulé, en fin de page

2 – Nettoyage des données :

2.1 – Suppression des adaptateurs

La suppression des adaptateurs s'effectue grâce à l'outil «cutadapt», disponible dans Galaxy sous le nom : «* Remove adaptators with cutadapt».

Lancer ce traitement sur le jeu s1 avec les paramètres suivants :

- Adaptateur : ATCTCGTATGCCGTCTTCTGCTTG
- Taille minimum : 18pb - Taille maximum : 25pb

Attention: miRdeep2 est un outil qui ne prend en entrée que des lectures comprises entre 18 et 25 pb.

Examiner les datasets de sortie.

2.2 – Statistiques après suppression des adaptateurs

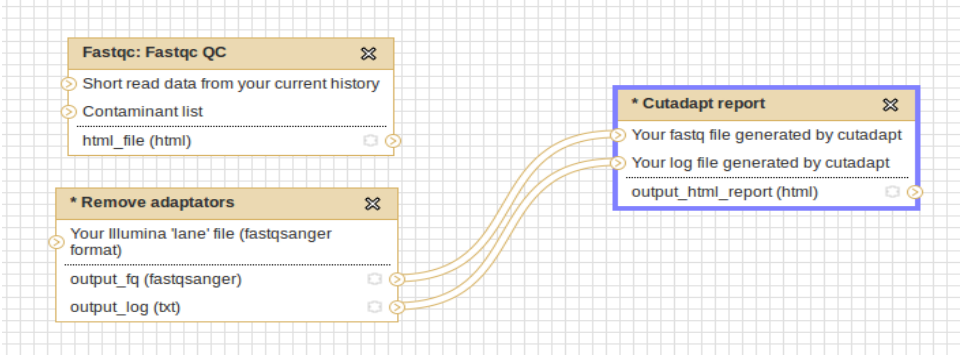
Lancer l'outil «*Cutadapt report after adaptators removing» sur les datasets du s1 pour produire des graphiques décrivant les données à l'issue de la suppression des adaptateurs.

Exemple de graphiques générés

Pour visualiser ces statistiques, ouvrir ou enregistrer le fichier sur votre PC et ouvrir la page «html» dans votre navigateur.

3 – Exporter votre historique en workflow

Créer un workflow sur la base des traitements précédents et l'exécuter sur le deuxieme fastq (s2). Le workflow est désormais utilisable pour nettoyer n'importe quel jeu de données sRNAseq. Penser à le renommer avec un nom parlant (ex : cleaning and QC of a sRNAseq FASTQ)



Il est possible de cacher certains datasets générés pendant le workflow et qui ne seront pas utilisés ultérieurement. Au moment d'éditer le workflow, l'astérix à côté de chaque output peut être cochée ou décochée.

Pour des workflows importants, cette possibilité prend tout son sens.

4 – Changer le format des FASTQ en FASTA

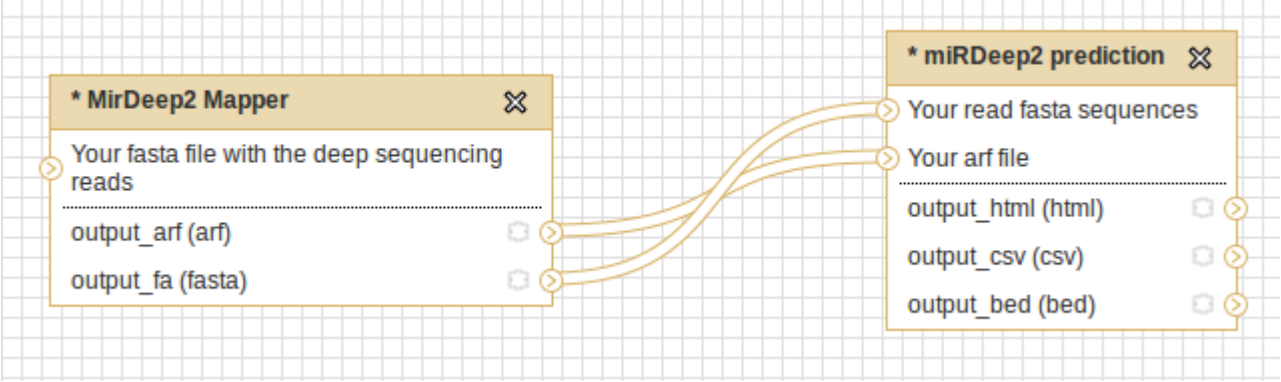
L'outil miRdeep2 disponible dans Galaxy a été wrappé de façon à ce que les fichiers d'entrée soient au format FASTA (les deux formats sont tolérés en ligne de commande).
Utiliser l'outil «FASTQ to FASTA converter» pour convertir les fichiers FASTQ.

Attention: prendre en input les fichiers issus du cutadapt.

Exercice n°3 : Recherche de miRNAs avec mirdeep2

MirDeep2 est un outil qui génère beaucoup de fichiers temporaires et qui nécessite beaucoup de temps de calcul. Dans le cadre de ce TP, utiliser le dataset présent dans l'historique «DATA TP 15/11/2013 sRNAseq UPS» : **cleaned_sequences.fasta**. Il s'agit d'un fichier au format FASTA contenant un nombre de séquences limité pour abréger les temps de traitement.

Le workflow dédié à la recherche de miRNA est disponible dans «Published Workflows» :
«**sRNAseq miRdeep2 workflow**».
Importer ce workflow et le lancer en prenant en entrée le fichier FASTA et le génome V4_454Scaffolds_filter.
La première brique de ce workflow permet de mapper les lectures contenues dans le fichier FASTA (en utilisant bowtie) sur la référence. La seconde brique permet d'effectuer une recherche (prédiction) de miRNA.



Exercice n°4 : Annotation des reads contre des banques de ncRNA

Quelques liens (outils / base de données) :

- BWA : http://bio-bwa.sourceforge.net
- BWA man : http://bio-bwa.sourceforge.net/bwa.shtml
- SAMtools : http://samtools.sourceforge.net
- mirBase : ftp://mirbase.org/pub/mirbase/CURRENT/ (hairpin.fa.gz)
- Rfam : ftp://ftp.sanger.ac.uk/pub/databases/Rfam/CURRENT/ (Rfam.fasta.gz)
- tRNA : http://gtrnadb.ucsc.edu/download.html (eukaryotic-tRNAs.fa.gz)
- rRNA : ftp://ftp.arb-silva.de/current/Exports/ ([LS]SURef_108_tax_silva_trunc.fasta.tgz)

Les banques de ncRNA sont déjà disponibles dans Galaxy.

L'objectif est de comparer les séquences annotées contre 3 banques différentes (miRbase_hairpin / Rfam / eukaryotic-tRNAs)

Construire un workflow permettant de générer un diagramme de Venn des annotations sur 3 banques différentes

Pour générer le diagramme de Venn, l'outil en question a besoin de 3 fichiers contenant les identifiants des séquences annotées contre chaque banque. Ces séquences doivent être triées dans le même ordre.

L'outil BWA prenant en entrée des fichiers au format FASTQ, il vous faut utiliser le dataset : **cleaned_sequences.fastq**.

Il est également intéressant de garder les informations concernant les annotations, notamment le nom de la lecture, son annotation et sa séquence.

Vous pouvez créer un workflow travaillant sur une banque à la fois ou bien créer un workflow traitant les 3 banques simultanément. Dans le premier cas, l'outil qui génère le diagramme de Venn ne sera pas intégré au workflow, contrairement au second cas.

Il est possible de rediriger le workflow vers un autre historique. Il faudra alors faire un transfert de datasets pour récupérer vos datasets importants pour la suite.

Étapes successives :

- Mapper les séquences contre une banque «Map With BWA for Illumina»
- Garder les séquences alignées (flags 0 ou 16) *(cf format SAM)* «Filter»
- Ne garder que les séquences ayant 0 ou 1 mismatch (expr : NM:i:[01]) «Select»
- Générer un fichier tabulé contenant le nom de la lecture, son annotation et sa séquence (c1,c3,c10) «Cut»
- Récupérer uniquement les identifiants des séquences «Cut»
- Trier ces identifiants de façon alphabétique dans l'ordre ascendant «Sort»
- Générer le diagramme de Venn «* Annotations comparison with Venn diagram»

Comparer l'effet du nombre de mismatchs sur les diagrammes de Venn

Relancer votre workflow en changeant le paramètre concernant le nombre de mismatchs.

Le nombre de mismatchs tolérés au moment de l'alignement a un impact non négligeable sur les séquences annotées, et donc sur le chevauchement des annotations.

Il faut donc être prudent lorsqu'on annote des séquences aussi petites et qu'un outil comme bwa tolère jusqu'à 3 mismatchs pour des séquences de cette taille (<25).

Exercice n°5 : Annotation des prédictions de miRdeep2 contre miRbase

En sortie, miRdeep2 génère 3 datasets :

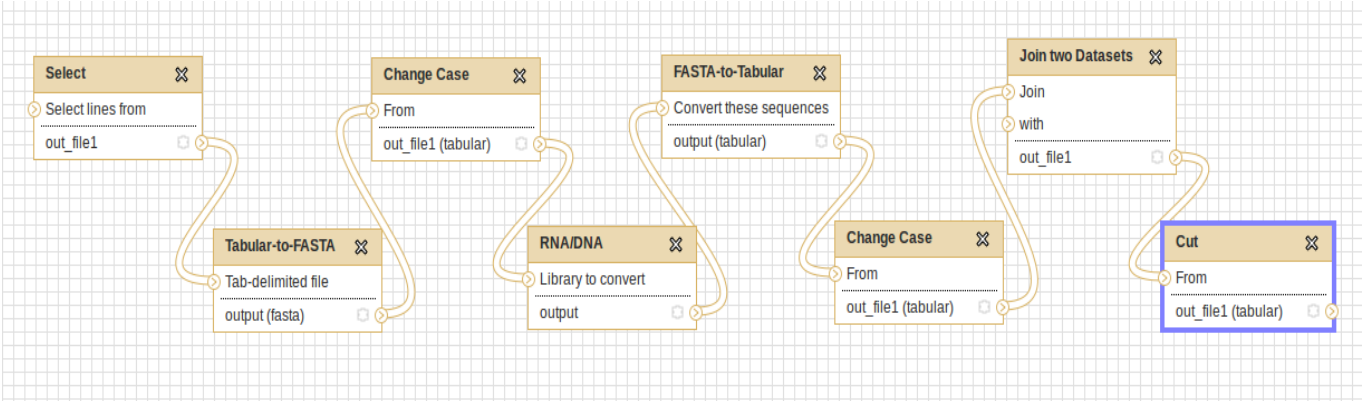
- Un premier fichier HTML dans lequel se trouvent les représentations graphiques des pré-miR, des repliements...
- Un second fichier BED dans lequel se trouvent les localisations des pré-miR
- Un troisième fichier CSV dans lequel se trouvent des informations de scores, d'annotations éventuelles sur des banques, et ce qui va nous intéresser, les informations de séquences.

Créer un workflow d'annotation des prédictions de miRdeep2

À partir de ce fichier CSV, créer un workflow capable de générer un fichier de prédictions contenant l'identifiant de la prédiction, la séquence nucléotidique du mature et l'annotation si elle existe. Si la prédiction n'a pas d'annotation, un «/» doit figurer.

Il doit donc y avoir dans votre fichier de sortie autant de prédictions qu'il y en avait au départ.

Descriptif du workflow :



Pour vous aider à comprendre ce que fait ce workflow, il est disponible : «sRNAseq annotation of miRdeep2 predictions»